Statistica Sinica **35** (2025), 1-22
doi:https://doi.org/10.5705/ss.202022.0108

SUBSAMPLING BASED COMMUNITY DETECTION FOR LARGE NETWORKS

Sayan Chakrabarty[1], Srijan Sengupta[2] and Yuguo Chen*[1]

[1] *University of Illinois Urbana-Champaign*
and [2] *North Carolina State University*

Abstract: Large networks are becoming pervasive in scientific applications. Statistical analysis of such large networks is prohibitive due to exorbitant runtime and high memory requirements. We propose a subsampling based divide-and-conquer algorithm, `SONNET`, for community detection in large networks. The algorithm splits the original network into multiple subnetworks with a common overlap, and carries out detection algorithm for each subnetwork. The results from individual subnetworks are aggregated using a label matching method to get the final community labels. This method saves both memory and computation costs significantly as one needs to store and process only the smaller subnetworks. This method is also parallelizable which makes it even faster.

Key words and phrases: Community detection, computational efficiency, degree corrected blockmodel, spectral clustering, stochastic blockmodel, subsampling.

1. Introduction

Network data appears in a wide variety of scientific and technological disciplines, such as social media (Sarkar and Rózemberczki, 2021), epidemiology (Leitch, Alexander and Sengupta, 2019), neuroscience (Roncal et al., 2013), and transportation (Gastner and Newman, 2006). A number of statistical models have been developed for analyzing such network data, starting with the homogeneous random graph model proposed by Erdös and Rényi (1959). In recent years, there has been substantial interest in blockmodels, such as the stochastic blockmodel (SBM), that allow nodes to be partitioned into different communities or blocks (Holland, Laskey and Leinhardt, 1983; Goldenberg et al., 2010). A number of generalizations of the SBM have been developed, such as the mixed membership blockmodel (Airoldi et al., 2008), the degree corrected blockmodel (DCBM) (Karrer and Newman, 2011), the popularity adjusted blockmodel (PABM) (Sengupta and Chen, 2018), etc.

One of the main inferential tasks on a network with an underlying community structure is to discover the community membership of each node. A number of community detection algorithms have been studied in the literature. This includes spectral clustering and its variants that leverage the eigen structure of the network

*Corresponding author. E-mail: yuguo@illinois.edu